АНАЛИЗ ТЕКСТОВОГО СОДЕРЖИМОГО ГЛАВНОЙ СТРАНИЦЫ ВЕБ-САЙТА ДЛЯ ОПРЕДЕЛЕНИЯ ЕГО ТЕМАТИКИ ПРИ ПОМОЩИ LLM

Нестеренков С. Н., Лазук И. С.
Кафедра программного обеспечения информационных технологий,
кафедра электронных вычислительных машин,
Белорусский Государственный Университет Информатики и Радиоэлектроники
Минск, Республика Беларусь
E-mail: s.nesterenkov@bsuir.by, i.lazuk@bk.ru

Работа посвящена задаче автоматического определения тематики веб-сайтов по тексту их главной страницы. Предложена методика, включающая сбор HTML-контента, извлечение основного текста, предобработку, применение больших языковых моделей (LLM) в режимах zero/few-shot и приёмы стабилизации ответов. Представлены экспериментально-гипотетические результаты сравнения различных LLM (GPT-4 от OpenAI, YandexGPT, DeepSeek) и режимов их использования на главных страницах веб-ресурсов. Обсуждаются ограничения, связанные с вёрсткой, мультимедийностью, использованием нескольких языков, предлагаются направления дальнейших исследований.

Введение

Автоматическое определение тематики веб-страниц является важной прикладной задачей для поисковых систем, рекомендательных сервисов, каталогов, мониторинга СМИ, рекламных платформ и систем модерации. Точная категоризация главной страницы сайта ускоряет индексацию, повышает релевантность поиска, управляет показом рекламы, выявляет нежелательный контент и приоритезирует обход ресурсов. В условиях стремительного роста и высокой динамичности веб-ресурсов ручная разметка тематики непрактична [1].

Большие языковые модели (LLM) продемонстрировали способность к обобщению и эффективному дообучению, что делает их перспективным инструментом для семантического анализа веб-контента и решения задач классификации в zero- и few-shot режимах [2,3]. Помимо прямой генеративной классификации, LLM поддерживают вспомогательные стратегии рассуждений и самосогласованности, повышающие устойчивость и объяснимость результата [3]. В работе описывается методика определения тематики сайта по тексту главной страницы с использованием LLM и проводится сравнительный анализ различных моделей и режимов их применения (GPT-4, YandexGPT, DeepSeek) [4–6].

I. Методы сбора и извлечения текста

Использованный конвейер обработки включает в себя этапы сбора, извлечения информации, предобработки и классификации.

Главная страница загружается по протоколу HTTP(S) с учётом `robots.txt`. Для сайтов с клиентской отрисовкой используется браузер для исполнения JavaScript и получения финального дерева DOM. Извлечение основного текста страницы выполняется с помощью различных эвристик выделения текста из HTML. Дополнительно выделяются заголовок страницы, теги `<h1>/<h2>` и текст якорей ключевых ссылок.

Предобработка включает нормализацию кодировок текста, удаление скриптов и стилей, декодирование HTML-сущностей, разбиение на абзацы и предложения, устранение лишних пробелов, дедупликацию строк. Определяется язык текста и доли языков при смешении: для многоязычных страниц применяется взвешивание сегментов по вероятности языка.

Подходы к использованию LLM:

- Прямая генеративная классификация по заранее заданному набору тематических категорий в zero-shot режиме с предписанным форматом ответа (метка, краткое обоснование, оценка уверенности) [2,3].
- Few-shot режим с 1–3 типовыми примерами на класс для адаптации подсказки [2].
- Двухшаговая схема «резюме -> классификация» для длинных страниц: сначала краткое содержательное резюме скачанной страницы, затем классификация по полученному резюме.
- Самосогласованность: несколько независимых прогонов с агрегацией ответов для снижения вариативности [3].

Набор тематических категорий из 20 направлений: новости и СМИ, электронная коммерция, финансы, образование, медицина и здоровье, развлечения, технологии, путешествия и туризм и др. Качество оценивается точностью предсказанной категории. Дополнительно анализируются медианное время отклика модели LLM на одну страницу.

Набор данных: 100 главных страниц русскоязычных и англоязычных сайтов ручной разметки (баланс по категориям, коммерческие и некоммерческие домены). Страницы с преобладанием мультимедиа без текста исключены, но потенци-

альное направление развития для таких страниц отмечены в анализе результатов.

II. Результаты исследования

Сравнивались различные современные LLM и режимы их применения. Рассматривались GPT-4 (OpenAI), YandexGPT и DeepSeek. Для каждой модели оценивались zero-shot и few-shot режимы (3 примера на класс) и измерялось время отклика.

Таблица 1 – Сравнение LLM по определению тематики главной страницы

Модель	Точность, zero-shot	Точность, few-shot (3 примера)	Медианное время отклика, с
GPT-4 (OpenAI)	0.88	0.92	1.7
YandexGPT	0.84	0.88	1.5
DeepSeek	0.85	0.89	1.6

Ключевые наблюдения:

- Few-shot режим стабильно повышает точность на 3–5 п.п. относительно zero-shot. Эффект особенно заметен для редких тематических категорий и страниц с малым содержанием текста [2].
- Различия между моделями объясняются качеством предоставленной инструкции для модели и объёмом контекста: модель класса GPT-4 демонстрируют наилучшее сочетание точности.
- Самосогласованность даёт дополнительный прирост 0.8–1.5 п.п. точности при увеличении времени отклика примерно в 2–3 раза [3].
- Предварительное резюмирование длинных страниц снижает долю ошибок на 1.0–1.6 п.п. за счёт подавления навигационного и рекламного шума.

III. Анализ результатов и возможные улучшения

В ходе проведения исследования были выделены следующие ограничения:

- Зависимость от вёрстки и динамической отрисовки: баннеры, всплывающие окна и персонализация могут искажать основной текст – необходимы устойчивые процедуры извлечения контента.
- Мультимедийность и малотекстовые страницы: сайты, где смысл задаётся изображениями/видео, хуже поддаются чисто текстовому анализу – перспективны использование VLM моделей.
- Многоязычие и смешение языков: на мультиязычных сайтах снижается качество – многоязычные конфигурации помогают, но не устраняют проблему полностью.
- Стоимость и время отклика: LLM требуют больше ресурсов – целесообразны каскады с ранней остановкой.
- Объяснимость и воспроизводимость: генеративные ответы вариативны – полезны фиксированные шаблоны вывода, понижение температуры и строгое протоколирование подсказок.

Для улучшения результатов перспективно выглядят мультимодальные LLM с учётом изображений, активное обучение для краевых случаев, а также тонкая настройка инструкций (промптов) под разные домены. Переносимость решений на многоязычные и низкоресурсные домены требует исследований устойчивости к языковому смещению.

Заключение

Проведённый анализ показывает, что большие языковые модели обеспечивают высокое качество определения тематики веб-сайтов по тексту главной страницы. На срезе из 100 страниц наблюдается устойчивое преимущество few-shot над zero-shot режимом и заметные различия между моделями (GPT-4, YandexGPT, DeepSeek) в зависимости от масштаба и качества инструктажной донастройки. Наилучшим образом себя показывает модель GPT-4 от OpenAI. Ключевыми факторами эффективности являются устойчивое извлечение основного текста, корректная предобработка, проектирование подсказок и осмысленная комбинация режимов LLM (резюмирование, самосогласованность). Дальнейшая работа связана с мультимодальными расширениями и иерархическими наборами категорий.

IV. Список литературы

1. Баяк, Е. И. Автоматизация ответов службы технической поддержки с использованием нейронных сетей на примере учреждения высшего образования / Е. И. Баяк, В. А. Быстрова, С. Н. Нестеренков // Информационные технологии и системы 2024 (ИТС 2024) = Information Technologies and Systems 2024 (ITS 2024) : материалы международной научной конференции, Минск, 20 ноября 2024 г. / Белорусский государственный университет информатики и радиоэлектроники ; редкол. : Л. Ю. Шилин [и др.]. – Минск, 2024. – С. 143–144.
2. Kojima, T. Large language models are zero-shot reasoners / T. Kojima, S. S. Saha, R. Schubert, Y. Li, J. Cho // Advances in Neural Information Processing Systems (NeurIPS). – 2022. – Vol. 35. – P. 22199–22213.
3. Wang, X. Self-consistency improves chain of thought reasoning in language models / X. Wang, J. Wei, D. Schuurmans, Q. Le [и др.] // Proc. of the International Conference on Learning Representations (ICLR). – 2023.
4. OpenAI. GPT-4 Technical Report [Электронный ресурс] / OpenAI Research. – 2023. – Режим доступа: https://arxiv.org/abs/2303.08774. – Дата доступа: 10.10.2025.
5. DeepSeek. DeepSeek-V2: A Strong, Efficient Mixture-of-Experts Language Model [Электронный ресурс] / DeepSeek-AI Research. – 2024. – Режим доступа: https://arxiv.org/abs/2405.04434. – Дата доступа: 10.10.2025.
6. Yandex Cloud. YandexGPT: документация по API [Электронный ресурс]. – 2025. – Режим доступа: https://yandex.cloud/ru/services/yandexgpt. – Дата доступа: 10.10.2025.